基于自然语言处理的期刊新媒体智能编作交互系统研发与应用

张芃捷　王　东　丰瑞兵　陈　健　毕　丽*
（重庆市卫生健康统计信息中心，重庆 401120）

摘　要：随着科技的不断进步，人与人之间的交流更加快捷方便。然后，由于作息时间交叉，受交流方式影响，作者与编辑、编务之间的沟通交流矛盾日益凸显，甚至成为限制期刊进一步发展的短板。在媒体融合发展背景下，如何引入人工智能，在第一时间解决作者的燃眉之急，又能保证编辑日常工作顺利进行，是值得研究的具体问题。本研究采用人工智能领域中重要的分支——自然语言处理相关技术，搭建了基于语义分类、相似度识别的新媒体智能编作交互系统。该系统可作为传统期刊投审稿系统与微信公众号的桥梁，不仅能实现智能问答、查稿等功能，及时解决作者常规问题和需求，解放编辑和编务的生产力，开展其他创新性期刊服务工作，同时盘活数据库中的作者、专家资源，引导其向期刊新媒体汇聚，为期刊融媒体发展奠定基础。

关键词： 人工智能；自然语言处理；新媒体；科技期刊　　**中图分类号：**G124　　**文献标识码：**A
文章编号：1671-0134（2021）12-146-03　　DOI：10.19483/j.cnki.11-4653/n.2021.12.047
本文著录格式：张芃捷，王东，丰瑞兵，陈健，毕丽．基于自然语言处理的期刊新媒体智能编作交互系统研发与应用 [J]. 中国传媒科技，2021（12）：146-148.

2015 年，国家新闻出版署颁布的《关于推动传统媒体和新兴媒体融合发展的指导意见》就如何推动传统媒体和新兴媒体融合发展提出指导意见。2020 年中共中央办公厅、国务院办公厅印发了《关于加快推进媒体深度融合的发展意见》，为学术期刊的媒体融合发展确定了重点，吹响了融合改革的“号角”。2020 年 12 月，中国科学技术信息研究所发布《2020 年中国科技论文统计结果》，结果显示，2010 年至 2020 年 10 月，中国科技人员发表国际论文 301.91 万篇，比 2019 年的统计结果同期提高了 15.8%；论文被引用 3605.71 万次，比 2019 年的统计结果提高了 26.7%。在科技论文产出不断提高的背景下，科技期刊的数字化转型与融合发展迎来新的机遇与挑战。一方面，以人工智能、新媒体为代表的新技术、新事物的引入，为出版领域注入活力，未来发展潜力让人充满遐想；另一方面，究竟在哪个地方引入与融合，具体效果如何，存在哪些问题，可以向哪些方向继续改进和延伸等问题缺乏具体案例经验，为科技期刊发展的布局和策略带来挑战。因此，以具体应用场景为切入点，将人工智能引入科技期刊的新媒体融合发展并研究其相关应用经验具有重要意义。

1. 已有研究分析

随着人工智能技术的不断发展，众多学者在人工智能与期刊、融媒体发展方面开展了研究。学者们提出，以用户为核心要素，采用人工智能算法对读者进行准确定位，能满足读者个性化需求。[1-2] 出版企业数字化转型中，利用人工智能进行学术出版流程再造，能实现传统出版产品和数字产品一体化、协同化、同步化，知识服务智能化等愿景。[3-4] 人工智能更多地会针对传统编辑行业当中的简单性工作及重复性工作，构建智慧出版模式。[5-6] 例如，中国大百科全书出版社与中科院合作研发人工智能产品“司南君”，能实现人机互答，丰富了出版业态 [7]；有医院针对医学检验仪器新入职维修人员因经验不足，设计了人工智能医学检验仪器故障智能问诊系统，能为维修人员提供可靠的维修建议。

这些实践探索为传统期刊与新媒体的进一步融合提供了思路与参考，但由上述分析可知，国内已有研究很少有期刊与新媒体具体融合案例和经验。本研究针对编辑 / 编务和作者交互（后简称编作交互）这个具体应用场景进行探索和尝试，通过引入人工智能分支技术之一——自然语言处理的相关技术，搭建连接投审稿系统与微信公众号的智能交互系统，以提高科技期刊运营效率。这是传统媒体与人工智能技术相融合的交叉性研究，也是学术期刊在新媒体领域的开创性探索研究。

2. 智能编作交互的需求

2.1 传统编作交互的矛盾

编作交互却逐渐成为限制期刊进一步发展的短板。一方面，当前编作交互主要为电话、投稿系统站内信息等传统交互方式，作者通常只能在工作时间拨打电话，且有时会出现占线、编辑 / 编务临时有事而无法接通咨询

基金项目：重庆市科学技术期刊编辑学会科研项目“基于自然语言处理的期刊新媒体智能编作交互系统研发与应用”（项目编号：CQKJQKXH2020003）；中国科学院自然科学期刊编辑研究会研究课题项目“智能化期刊投审稿系统自然语言处理模块的应用探索”（项目编号：YJH202109）。

（* 为本文通信作者）

的情况，编辑/编务很难在第一时间解决作者的燃眉之急；而作者通过投稿系统发送站内信息，存在较严重的信息滞后性，交互效率低下，且需要作者通过 PC 电脑发送站内信息，操作复杂。另一方面，以重庆市卫生健康统计信息中心下属期刊《现代医药卫生》为例，作者大多数咨询内容为期刊情况、投审稿流程等投稿前常规问题，或者是稿件刊登、发票和快递情况等投稿后相关问题，前者内容重复率高，后者需要编辑/编务查询系统后回复，耗费时间长，加之来电数量大，导致相关编辑/编务除接电话和回复站内信息外，很难开展其他创新性工作，工作时间被严重挤压，人力资源浪费情况严重。然而，编作交互又是期刊出版过程中的重要环节，在期刊守正创新、提升作者服务质量和期刊声誉方面的作用不可忽视，需要期刊花大力气做好该项工作。随着未来期刊的投稿量不断上升，期刊融合化、数字化转型的进一步深入，编作交互矛盾对期刊的影响正不断放大。

2.2 “沉睡”的投审稿系统数据库资源

乘着移动互联网络、智能终端及软硬件技术蓬勃发展的东风，新媒体应声而起。众多期刊媒体也开始以新媒体为基础进行融媒体探索，尝试建立和运营微信公众号，但都遇到类似的问题：缺乏流量和用户基础。根据西瓜数据发布的《2020 年公众号生态趋势调查报告》，微信公众号创作者数量已超 2000 万，全网流量竞争日趋激烈。对于科技期刊的微信公众号而言，即便是花钱引流也很难获得目标用户，长此以往将逐步失去活力甚至被边缘化。另一方面，科技期刊的投审稿系统数据库中“沉睡”着众多宝贵的数据资源——忠实于期刊的作者、专家数据。如果能通过有效手段复用这些资源，逐步通过实用功能和方法将作者、专家引向微信公众号，将使后者获得长足发展，为期刊的数字化、融媒体转型打下基础。[8]

2.3 编作交互的需求

人工智能的加入，推动了编作交互流程的改进，使其最大限度地满足作者、专家和编辑三方需求。

编辑在岗在位的时候，作者、专家仍然可以通过电话等常规方式联系编辑同步处理较为紧急的棘手问题；另外，可以通过人工智能的系统回答重复性的咨询内容，满足作者个性化查询，尽可能地随时解决作者的困惑与需求。在此基础上，向数据库中的作者、专家进行提示与引流，逐步向微信公众号迁移作者、专家资源。加上期刊自身在微信公众号上的内容开发，最终实现微信公众号的跨越式进步。

3. 智能编作交互系统研发与应用

参考上述的编作交互需求，本研究将研发和应用分为交互功能研发、系统与数据库通信连接、系统与微信公众号通信连接、整合完善等四个部分。

3.1 智能编作交互系统的基本架构

由于较多期刊的投审稿系统是在 2010 年前后建立，彼时社交网络刚刚萌芽，微信也没有公众号一说，故投审稿系统大多为独立封闭系统，并没有延伸功能或相关接口。在此情况下，本研究考虑研发独立运行的系统，既连接微信公众号，又不会对原有的投审稿系统和工作流程产生影响。

智能编作交互系统类似一座桥梁，沟通微信公众号与投审稿系统数据库。首先，系统在微信公众号部分获取来自用户的文字咨询，将其传入意图识别模块进行识别。如果判定为常规问答，则通过相似文本拟合找寻最接近的问题并返回答案；如果判定为查询，则将文字转化为 SQL 数据库查询指令，通过投审稿系统数据库获取稿件的审稿情况，并根据稿件状态给予作者对应的建议。最终，系统再将结果以文字回复用户，实现随时随地的交互流程（如图 1）。

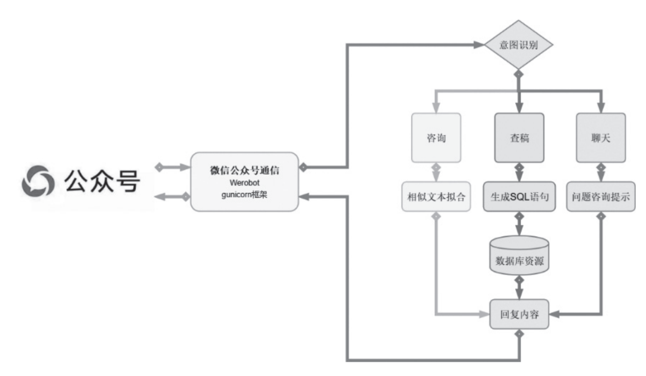


图 1 智能编作交互系统运行流程

3.2 核心功能研发

智能编作交互系统的核心功能包括 3 个部分：意图识别、相似文本拟合与数据库查询。

3.2.1 基础准备

作为人工智能的重要分支，自然语言处理主要实现人与计算机之间利用自然语言进行有效通信的各种方法。其中，如何用数学语言来表示文本，继而将其转化为实现某项功能的模型，是建立系统的基础。本研究以词向量 Word2vec 将单个文字转化为可供计算数字单位，并利用 Fasttext 思路[9]，将句中各单字向量相加后求平均值，得到该句话的平均向量。两者的精度高、优化效率高，能有效提高分类准确率和相似文本匹配率。

编程方面采用 Python 语言编程，主要用 Jupyter Notebook 进行实时编码与调试。以重庆市卫生健康统计信息中心下属期刊《现代医药卫生》投审稿系统数据库 2020 年及以前的 50000 余条投审稿数据（包括稿件状态、作者信息、所属编辑等）作为查询基础，待后期进行数据的应用。

3.2.2 意图识别

意图识别是对输入的问询内容进行分类处理。其会调用已保存好的文字表征矩阵和支持向量机（SVM）分类模型，将作者经微信公众号的问询内容转为平均向量，输入 SVM 分类模型进行预测，获得意图识别结果。结果分为 3 种：“咨询”，转入智能问答流程；“查稿”，转入数据库查询流程；“聊天”，则提示“AI 暂时不会侃大山”。

3.2.3 智能问答

智能问答流程采用相似文本拟合，其基础原理是计算两个句子向量夹角的余弦值（余弦相似度），用于衡量两个句子之间的相似性。当系统将问询内容转为平均向量后，利用 Sklearn 工具库的 Cosine_similarity 函数，对

比问询内容和备选问题的余弦相似度，选取与问询内容最接近的备选问题，然后将对应问题的回复返回给作者。

3.2.4 数据库查询

转入数据库查询流程后，系统会扫描整个问询句子，利用正则表达式获取稿号（《现代医药卫生》是以“S”+10位数字组成）或作者名，然后引入 Pymssql 工具库，于 Python 程序端执行 Sql 语句，在数据库中进行查询和匹配，成功后提取数据库中稿件的相关信息。组合定式文，回复作者稿件所处状态，以及下一步将要开展的工作。

3.3 与微信公众号通信连接

无论是订阅号还是服务号，是回复信息交流还是发布推文，独立于微信公众号原网页进行的操作都需要进行微信公众号开发。其中，主要是移动端网页的页面开发，与微信公众号通信的关键则是搭建合理的开发框架，完成微信公众号与智能编作交互系统的通信。

由于服务器为CentOS7，且考虑要整体服务需要简单、快速，本研究采用了 Gunicorn Python WSGI HTTP Server 作为微信公众号开发框架；引入 Werobot 工具库，解析微信服务器发来的信息数据，从中提取作者发送的消息内容进行计算，并将回复内容打包成可识别的信息数据传回微信服务器。

3.4 系统的整合完善

智能编作交互系统包含两个部分，训练部分与应用部分。两个部分可以同步进行，异步迭代升级。

在训练部分，首先纳入所有训练数据集文本，去掉里面的停止词（Stop words，表示实际语言意义的字词）后，拆分为单个文字，纳入 Gensim 工具库进行训练，获得 Word2vec 的文字表征矩阵，输入拆分好的数据集文本句子，转化为向量后，利用 Fasttext 计算该句话的平均向量。将所有训练数据集文本的平均向量值输入 Sklearn 工具库，训练支持向量机（SVM）分类模型（技术成熟，训练速度快，精度高）。最后，将文字表征矩阵和 SVM 分类模型保存，待意图识别的应用部分调用。

应用程序启动以后会永续运行，检测用户发给微信公众号的内容，及时进行计算与回复。由于将训练与应用分开，即使有文字内容更新要进行训练或者训练模块出现问题，也不会对正在进行的应用程序造成影响，后者会继续采用已有的模型进行计算和回复。完成训练后，重启应用程序即可实现内容的迭代升级。

3.5 有待解决的问题

在测试用个人公众号（公众号名为“村长 NLP 自留地”）投入使用后，智能编作交互系统已经能识别问题并进行针对性回答。在研发过程中，本研究仍也有一些问题尚待解决和完善。本研究中导入的文本数量有限，覆盖问题的方向和范围不足，采用模型和算法以效率为优先目标，有时候容易出现“答非所问”的现象，有待进一步扩充文本内容，优化算法与模型，同时建立行之有效的罕见问题收集、整理机制，最终提高作者咨询的满意度。此外，受限于泛化能力优先，目前“聊天”功能尚未开放，未来可针对期刊所面向的专门领域，研发可供作者了解该领域相关知识点、研究进展的人工智能产品，实现人机互答聊天。

4. 小结与展望

通过对自然语言处理技术的实际探索应用发现，系统可以通过已有的文字、数据进行学习、归纳，完成需求内容的预测[10]，搭建与传统投审稿系统的通信桥梁，盘活数据库中的数据资源。可以预见的是，以此为基础，继续对智能编作交互系统的相关功能进行扩充和延伸，可能实现基于微信端的审稿、定向（对某位编辑）咨询、投审稿全流程提醒与服务支持等功能。如果将智能编作交互系统作为单一功能模块纳入期刊数字化融媒体平台，并同步加入会议服务[11]、知识服务[12]、在线培训等功能模块，并深度与投审稿系统进行对接开发，实现功能协同，其成品或将成为期刊数字化转型和融媒体发展的重要思路和方向。媒

参考文献

[1] 刘焕英．疫情下人工智能与科技期刊融合发展探析 [J]. 出版广角，2020（7）：26–28.

[2] 江雨莲，孙澂．人工智能在医学期刊编辑出版中的应用 [J]. 科技与出版，2020（2）：66–71.

[3] 李媛．人工智能时代的学术期刊数字化传播 [J]. 中国科技期刊研究，2019（11）：1183–1190.

[4] 刘华东，马维娜，张新新．“出版＋人工智能”：智能出版流程再造 [J]. 出版广角，2018（10）：14–16.

[5] 张勇，王春燕，王希营．人工智能与学术期刊编辑出版的未来 [J]. 中国编辑，2019（4）：64–68.

[6] 刘平，杨志辉．人工智能构建科技期刊智慧出版模式 [J]. 中国科技期刊研究，2019（5）：462–468.

[7] 范军，陈川 .AI 出版：新一代人工智能在出版行业的融合创新 [J]. 中国编辑，2019（5）：64–71.

[8] 周正浩，余夏琳，杨晓云．探析人工智能对传媒业的影响 [J]. 中国传媒科技，2020（10）：41–43.

[9]Ivan Nikolaev，Dmitry Botov，Yuri Dmitrin，et al. Use of Topic Modelling for Improvement of Quality in the Task of Semantic Search of Educational Courses[A]. Proceedings of the 21st International Workshop on Computer Science and Information Technologies [C]. Ufa： CSIT，2019.

[10][加] 阿杰伊·阿格拉沃尔，乔舒亚·甘斯，阿维·戈德法布，等 .AI 极简经济学 [M]. 长沙：湖南科技出版社，2018：8–20.

[11] 方琍．智能会议管理的设计方案 [J]. 数字化用户，2019（44）：101–102.

[12] 袁阳，肖洪．基于知识元库自动编辑的知识服务优化 [J]. 科技与出版，2017（6）：22–25.

作者简介：张芃捷（1987–），男，重庆，出版中级（编辑），研究方向：出版与人工智能融合发展。

（责任编辑：张晓婧）